Assessing the coverage of data collection campaigns on Twitter: A case study

Vassilis Plachouras[1] Yannis Stavrakas[2] Athanasios Andreou

Institute for the Management of Information Systems (IMIS)
ATHENA Research Center
Artemidos 6 & Epidavrou, Maroussi 15125, Athens, Greece
{vplachouras, yannis}@imis.athena-innovation.gr,
athan.andreou@gmail.com

Abstract. Online social networks provide a unique opportunity to access and analyze the reactions of people as real-world events unfold. The quality of any analysis task, however, depends on the appropriateness and quality of the collected data. Hence, given the spontaneous nature of user-generated content, as well as the high speed and large volume of data, it is important to carefully define a data-collection campaign about a topic or an event, in order to maximize its coverage (recall). Motivated by the development of a social-network data management platform, in this work we evaluate the coverage of data collection campaigns on Twitter. Using an adaptive language model, we estimate the coverage of a campaign with respect to the total number of relevant tweets. Our findings support the development of adaptive methods to account for unexpected real-world developments, and hence, to increase the recall of the data collection processes.

Keywords. Social networks, data management, event tracking

1 Introduction

There is a growing number of applications that analyze data from online social networks and microblogging platforms, in order to detect breaking news as they happen, or to monitor the interests of users with respect to ongoing events. The quality of the analysis results depends to a great extent on the availability of exactly those data that are relevant to the task at hand. Microblogging platforms such as Twitter, however, limit the access to the full stream of data, and applications typically employ one of the following alternatives for collecting thematically focused tweets: (a) tracking a number of terms or users over a small random sample of the full stream, or (b) executing repetitive queries of limited expressiveness that usually include a number of terms or users against which the full stream of tweets is matched.

[1] Supported by the EU/Greece funded KRIPIS: MEDA project
[2] Supported by the European Commission under ARCOMEM (ICT 270239)

While the first alternative may provide a balanced distribution of terms and users with respect to the full stream, allowing therefore the discovery of trending topics, it only returns a small percentage of the relevant content. On the other hand, the second alternative returns all the content matching the query conditions, but suffers when a topic evolves over time and the query conditions become gradually irrelevant. Consider the case of unexpected developments in an event, for example the terrorist attack during the Boston Marathon on April 15, 2013. Querying with an immutable set of hashtags (such as *#bostonmarathon*) would result in a significantly lower coverage of the relevant data, as the focus shifted from the marathon itself to the attack and the events that followed. The ideal approach would be to use the second alternative, but with varying query conditions that reflect the evolution of the topic in question.

In this work, (a) we describe a methodology to evaluate the coverage of data collection campaigns from Twitter on a given topic, based on the work by Lin *et al.* [1], and (b) we use this methodology to demonstrate the impact of unexpected developments during an event on the quality of the collected data. Specifically, we employ our evaluation framework to assess the coverage of four data collection campaigns about the 2012 and 2013 Boston marathons and the 2012 and 2013 London marathons. Our results show that the coverage of a data collection campaign achieved by querying with a topic specific hashtag is relatively high in case an event develops as expected. On the other hand, a data collection campaign based on a predefined immutable hashtag achieves a low coverage when the relevant event develops in unexpected ways, because users are more likely to spontaneously start using new hashtags reflecting the current developments. In the case of 2013 Boston marathon, where a terrorist attack took place, a data collection campaign based only on the hashtag *#bostonmarathon* would retrieve approximately 45% of the relevant tweets. Notice that adapting the search to the unexpected developments within a topic is of paramount importance in many applications, like for example in data-driven journalism. Consequently, in such cases, it is important to consider methods for the automatic adaptive updating of the query used to guide the data collection campaign.

The remainder of this paper is organized as follows. In Section 2, we describe a platform for the management of social network data we are developing and outline the motivation for this work. In Section 3, we present the methodology for evaluating the coverage of a data collection campaign. In Section 4 we describe the experimental setting and present the results. We discuss related works in Section 5 and we close with some concluding remarks in Section 6.

2 Collecting tweets with TwitHoard

Our need to assess the coverage of a data collection campaign from Twitter arises in the context of developing TwitHoard, a platform for collecting data from Twitter [3], and modeling the dynamics of terms and term associations [2].

The aim of TwitHoard is to aid users in defining and managing data collection campaigns on Twitter. The platform allows the concurrent running of multiple campaigns. A user defines a campaign by providing its duration, a set of terms or

hashtags, a set of Twitter usernames, and possibly specifying other filters on language or geographic locations. A campaign can be paused, restarted, and refined, by updating its definition at any time instance. The user can also enable the crawling of Web pages linked from tweets to archive them for future reference. Figure 1 shows a screenshot of TwitHoard's filtering screen, where a user can create a selection of tweets coming from a specific user, or containing a given hashtag.

TwitHoard will also integrate a model of the temporal evolution of entities and a set of query operators to enable users to create views of the collected datasets according to complex temporal conditions. For example, a journalist may be interested in finding the tweets during the period in which the association strength between the hashtags *#boston* and *#marathon* is increasing. The model and the query operators allow the expression of such queries with varying time granularities.

An important feature that will be integrated in TwitHoard, is the capability of the system to automatically adapt the campaign definition so as to reflect the evolution in the topic at hand. The study in this paper is a first step towards this direction.

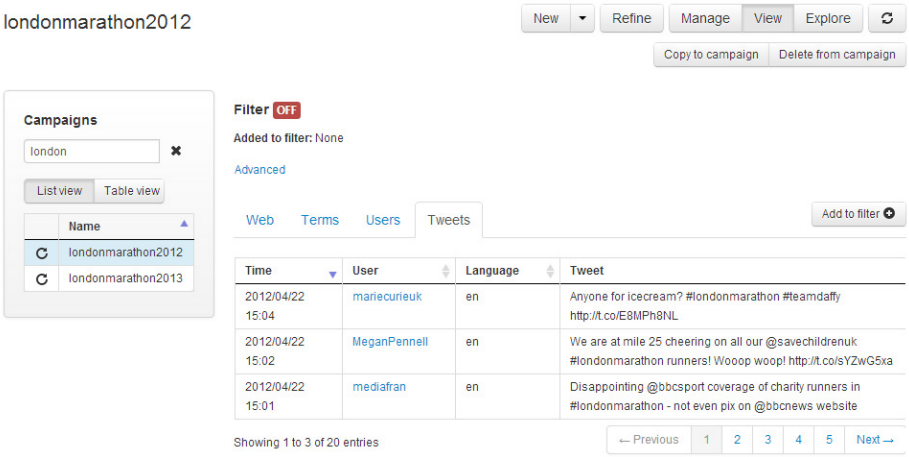

Fig. 1. Screenshot of TwitHoard campaign management platform from the 2012 London marathon dataset.

3 Methodology

In this section, we describe a framework for assessing the coverage of a data collection campaign. The framework is based on the work of Lin *et al.* [1], but our objective is different, as we explain in Section 5.

We assume that a data collection campaign is defined by a set of hashtags $H = \{h_1, h_2, \ldots, h_n\}$. Given a stream of tweets T, if the current tweet tw contains at least one hashtag from H, we add it to the ground-truth set G_w, where w is the maximum cardinality of the set. If adding a tweet in the ground-truth set increases its cardinality over

w we remove the oldest tweet to maintain the maximum cardinality equal to w. If the tweet tw does not contain any hashtag from H, then we classify it as *missed-relevant* or *non-relevant* to the campaign. Our objective is to estimate the number of tweets that are relevant to a campaign, but do not match any of the hashtags in H. We expect few such tweets for a campaign that is well characterized by the hashtags in H. However, for a campaign defined by a set of hashtags that do not capture well the topic or event of interest, or for a campaign about an event where there are unexpected developments, we expect that the number of relevant tweets that do not contain hashtags in H will be higher.

The relevance of a tweet to the campaign defined by H is determined by its similarity to the set G_w. More specifically, a tweet is relevant to the campaign if its perplexity with respect to the language model of the tweets in G_w is lower than a threshold k. In the remainder of this section, we describe the language model built from tweets in G_w and the perplexity classifier.

3.1 Adaptive language model for topic tracking

Given the set G_w of the w most recent tweets that match the campaign definition, we build a foreground language model, which is combined with a background model using Jelinek-Mercer smoothing [1]. In preliminary experiments, we have also used smoothing techniques based on absolute discounting, Dirichlet priors and stupid backoff, but Jelinek-Mercer smoothing resulted in models with lower average perplexity. The probability of a word x is given by the following equation:

$$P(x) = \lambda \frac{c(x, G_w)}{\sum_x c(x, G_w)} + (1-\lambda) P_B(x) \tag{1}$$

where λ is a hyper-parameter, $c(x, G_w)$ is the frequency of word x in the ground-truth set G_w, and $P_B(x)$ is the probability of x in the background model. As described in [1], the background model is also smoothed using absolute discounting with δ=0.5. We set the value of λ for Jelinek-Mercer smoothing such that we minimize the average perplexity of tweets containing at least one hashtag from H with respect to the set G_w. In other words, the value of λ is set such that we optimize the prediction of the next tweet by the language model $P(x)$.

3.2 Perplexity classifier

We use a simple perplexity classifier to decide whether a tweet, which does not contain any of the hashtags in H, is relevant to the data collection campaign defined by H, and thus, it would have been beneficial to also collect it. The perplexity of a tweet with respect to language model $P(x)$ is defined as follows:

$$pow\left(2, -\frac{1}{N}\sum_{i=1}^{N} \log_2 P(x_i)\right) \tag{2}$$

where N is the number of words in the considered tweet and x_i is the i-th word of the tweet. Perplexity expresses the surprise of seeing a sample of size N given the distribution $P(x)$. Hence, a lower perplexity value means that the considered tweet is more similar to the distribution $P(x)$. In our experiments, we mark a tweet as relevant if the computed perplexity is lower than a threshold k.

3.3 Assessing the coverage of a campaign

We evaluate the coverage of a data collection campaign in terms of the collected tweets that contain at least one of the hashtags from the campaign definition and the missed-relevant tweets, which are relevant but do not contain any of the specified hashtags. We use recall at time t, denoted by $R(t)$, which is defined as follows:

$$R(t) = \frac{G_\infty(t)}{G_\infty(t) + Missed(t)} \tag{3}$$

where $G_\infty(t)$ is the total number of tweets that belonged to G_w at some point in the past up to time t, and $Missed(t)$ is the number of missed-relevant tweets, i.e. the tweets encountered up to time t that are on topic but do not contain any hashtag from set H. A high value of recall $R(t)$ means that the definition of the data collection campaign captures well the topic or event of interest, and Twitter users writing relevant tweets are very likely to use at least one hashtag from H. On the other hand, a low value of recall means that there are many tweets, which are similar to the ones containing a hashtag from the campaign definition but which do not themselves contain any such hashtag. This may be either due to an incomplete definition of the campaign, or unexpected developments in the topic or event of interest, resulting in a change of the vocabulary present in relevant tweets.

4 Evaluation

We employ the methodology described in Section 3 to evaluate the coverage of four data collection campaigns about marathon events in Boston and London for 2012 and 2013. While there were no major incidents during either versions of the marathon in London and the 2012 Boston marathon, the 2013 Boston marathon was marked by the explosion of two bombs near the finish line, followed by the identification and the hunt of the terrorists. For each of the four events, we simulate a data collection campaign. The simulation targets a corpus of tweets that we have collected during the relevant time periods, by archiving a random sample of the Twitter stream. Our goal is to evaluate the coverage of each campaign, expecting that the coverage for the 2013 Boston marathon will be significantly lower than that of the rest of the campaigns.

4.1 Datasets

We have used Twitter's Streaming API to collect a set of tweets for each of the 2012 and 2013 Boston and London marathons. More specifically, we have collected

tweets containing at least one English stop-word. Given the high volume of tweets matching this condition and the limitations of the Streaming API, we have collected 50 tweets per second, and hence, more than four million tweets per day. We filter out tweets that do not contain any hashtag, after tokenizing the text of each tweet and removing stop-words; we further filter out those tweets that have fewer than 5 distinct words or fewer than 10 words in total. We run the simulated campaigns on these datasets using the hashtags shown in Table 1.

Table 1 shows the start and end times of each dataset, the number of tweets contained in each dataset after the filtering described above, and the set of hashtags H used to define the corresponding data collection campaigns. For the datasets covering the 2012 and 2013 London marathons and the 2012 Boston marathon, the data covers three days, starting the day before the event and ending the day after. The dataset for the 2013 Boston marathon covers 8 days in order to include tweets about the events that took place during and after the marathon.

Table 1. Description of datasets for the 2012 and 2013 London and Boston marathons. Start and end times are given in UTC.

Dataset	Start date	End date	# tweets	Hashtags H
London12	2012/4/21 00:00	2012/4/24 00:00	445,137	#londonmarathon, #vlm, #vlm2012, #bbcmarathon
London13	2013/4/20 00:00	2013/4/23 00:00	485,701	#londonmarathon, #vlm, #vlm2013, #bbcmarathon
Boston12	2012/4/15 00:00	2012/4/18 00:00	457,373	#bostonmarathon
Boston13	2013/4/14 00:00	2013/4/22 00:00	1,344,471	#bostonmarathon

4.2 Estimating coverage

We estimate the coverage of a data collection campaign as follows. For each tweet in the stream of tweets we have collected, we check whether it contains a hashtag from the set H. If this is true, then we add it to the set G_w and we update the language model $P(x)$. Otherwise we compute its perplexity with respect to the set G_w. If the computed perplexity is lower than the threshold k we mark the tweet as missed-relevant, because the data collection campaign would not download it.

We set the value of the λ hyper-parameter in Jelinek-Mercer smoothing from Eq. (1) as follows. For each dataset, we compute the average perplexity of tweets containing a hashtag in H with respect to G_w of the w most recently seen tweets containing a hashtag in H. We have used several values for w ranging from 100 to 10000. According to the results, $\lambda = 0.1$ was the value that most often resulted in the minimum average perplexity. For this reason, we fix $\lambda = 0.1$ for the remaining of the experiments. The background model $P_B(x)$ is built from a set of tweets collected during a period of one week from 2012/3/25 to 2012/3/31. The perplexity of the unigram language

model built from tweets in the ground-truth for Boston2013 is 252.84. The average perplexity of tweets in ground-truth for Boston2013 is significantly higher, because it is computed as new tweets are added to the ground-truth, and hence, the probabilities for previously unseen words are low.

We set the perplexity threshold $k = 5000$ for the classifier, meaning that a tweet is marked as relevant if it does not contain a hashtag in H and its perplexity with respect to the language model $P(x)$ is lower than 5000. The value of threshold k controls the trade-off between precision and recall. For example, a higher value of k will increase the number of missed-relevant tweets but they may not be on the same topic as tweets in G_w. The magnitude of the threshold k depends on the specific dataset we use. We have selected the value 5000 after performing preliminary experiments, where we observed a high number of relevant tweets. More specifically, for the Boston2013 dataset, we have manually inspected a random sample of 200 missed-relevant tweets and found that 87% were about the Boston marathon and subsequent events.

Table 2. Recall achieved for the data collection campaigns regarding the 2012 and 2013 London and Boston marathons.

Dataset	G_∞	Missed	Recall
London12	873	92	0.9047
London13	1105	99	0.9178
Boston12	245	10	0.9608
Boston13	6726	8255	**0.4490**

4.3 Results

Table 2 shows the achieved recall for the employed datasets after processing all available tweets. We set the parameter w to a high enough number so that practically G_∞ in Table 2 denotes the number of tweets containing any hashtag in H for each simulated campaign. The results in the table show that collecting tweets containing hashtags in H for the 2012 and 2013 London marathons and the 2012 Boston marathon retrieves the majority of the relevant tweets. For example, both campaigns about the London marathon would have gathered more than 90% of the relevant tweets. We attribute the high recall value to the fact that most of the tweets that are about the 2012 and 2013 London marathons do have one of the hashtags in the set H. Moreover, there was no unexpected development during either of the marathons in order to lead people to change the vocabulary used in their tweets. We obtain similar results for the 2012 Boston marathon. On the other hand, the recall for the 2013 Boston marathon is significantly lower, reaching only 0.4490. The low recall value is due to the terrorist attacks that took place at the 2013 Boston marathon and the following events. These unexpected developments have led users to employ other hashtags, such as *#prayforboston*, *#watertown*, *#boston*, in addition to or in place of *#bostonmarathon*. Next, we focus on the analysis of the results for the 2013 Boston marathon.

Figure 2 illustrates the temporal evolution of the hourly frequency of tweets in the ground-truth (containing the hashtag *#bostonmarathon*), and the missed-relevant

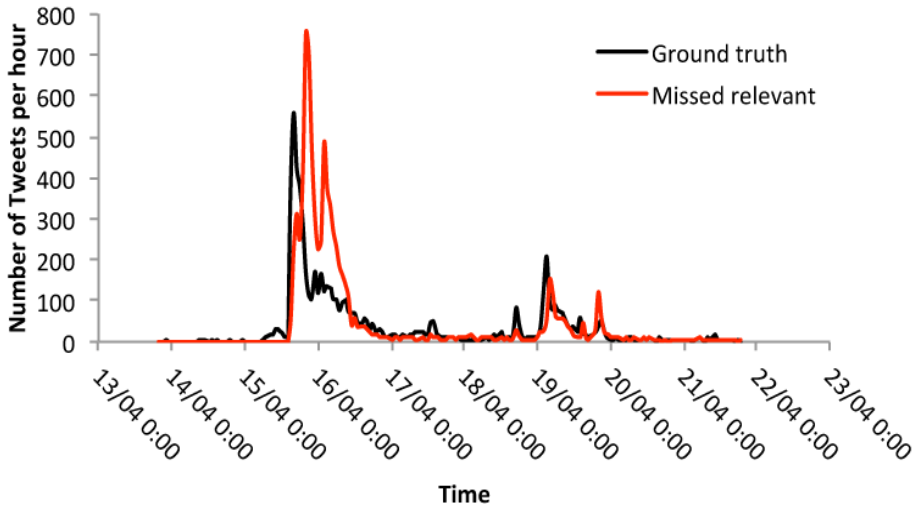

Figure 2. Hourly frequency of ground-truth and missed-relevant tweets.

ones, which are marked as relevant by the perplexity classifier but do not contain #*bostonmarathon* (times are shown in EDT). We observe that there is a significant spike in the use of #*bostonmarathon* at 15/04 15:00, just after the bomb explosions near the finish line of the Boston marathon. This spike is followed by a higher spike in the number of relevant tweets that do not contain #*bostonmarathon*. Similarly, we observe smaller spikes at 18/04 17:00 when images of the suspects are released, 19/04 2:00 when police hunts the suspects, and 19/04 20:00 when the second suspect is arrested. Each spike in the number of ground-truth tweets is followed by the number of relevant tweets that do not contain #*bostonmarathon*. Overall, we observe that a campaign that collects tweets relevant to the Boston marathon would need to adapt very quickly in order to increase the number of collected relevant tweets.

Table 3. Top-5 frequent hashtags in ground-truth and missed-relevant tweets.

Ground-truth		**Missed relevant**	
Hashtag	**Freq**	**Hashtag**	**Freq**
#prayforboston	701	*#prayforboston*	5564
#boston	342	*#boston*	1055
#watertown	152	*#watertown*	254
#fbi	121	*#breaking*	125
#breaking	78	*#marathon*	103

Next, we examine the frequency distribution of hashtags in the ground-truth and in the set of missed tweets. Table 3 shows the top-5 most frequent hashtags in the ground-truth set (excluding #*bostonmarathon*) and in the set of missed relevant tweets, respectively. We observe that four hashtags are common in both sets. Hence, we expect that by automatically updating the campaign definition to include some of the frequently co-occurring hashtags, we could collect more relevant tweets. Figure 3 shows the temporal evolution of the frequency of #*prayforboston* (top) and #*watertown* (bottom) when they occur in the ground-truth and in the missed-relevant tweets, respectively. We observe that for #*prayforboston* the start of the peak in missed-relevant tweets coincides with a smaller peak in the number of ground-truth tweets containing #*prayforboston.* For the hashtag #watertown, a peak in missed-relevant tweets is preceded by a smaller peak for the same hashtag in the ground-truth.

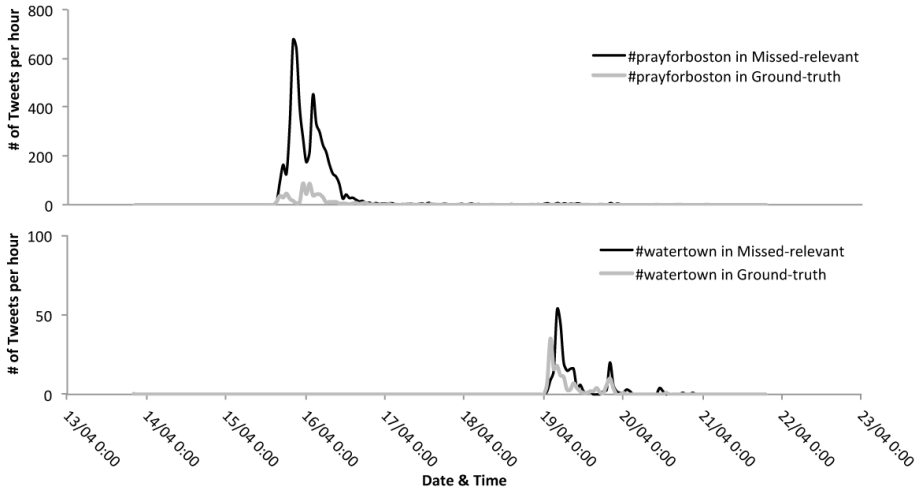

Fig. 3. Hourly frequency of tweets containing the hashtags #prayforboston and #watertown in the ground-truth and missed-relevant tweets, respectively.

5 Related works

Our work is broadly related to Topic Detection and Tracking (TDT) [4]. However, there are important differences. For example, we need to track the evolution of a topic or event by actively querying a social network API to obtain data. Hence, the query affects the data that is available for further processing.

The framework we have described is based on the work of Lin *et al.* [1], which studies various smoothing and history retention methods for adaptive language models built from tweets containing a given hashtag. They evaluate the adaptive language models in terms of perplexity and precision/recall, where the relevance of each tweet

is based on the output of a perplexity classifier. Our work, however, is different, because we consider all tweets that contain a given hashtag as relevant, and we estimate the number of tweets which would be relevant but do not contain the given hashtag.

Ward [6] presents an automatic query expansion method to collect tweets about TV programs. We do not make any assumption about the events or topics and, hence, we cannot use domain-specific knowledge as in [6] or other works for Social TV [5].

There is a number of works aiming to predict the popularity of hashtags in the future. Ma et al. [7] develop classifiers using content and contextual features to predict the popularity of a hashtag on a daily basis. Tsur and Rappoport [8] investigate content and temporal features to predict the popularity of hashtags with linear regression.

6 Conclusions

In this work, we have described a framework based on language models to assess the coverage of a data collection campaign on Twitter. We have employed this framework to evaluate the coverage of four simulated data-collection campaigns. Our results show that we can achieve high coverage of relevant data if the focus of the topic does not change significantly over time. However, if there are unexpected developments and users modify the vocabulary of their status updates, the coverage is harmed considerably. In our use case of marathon races, this finding is observed for the 2013 Boston marathon, which was marked by the explosion of two bombs. Overall, our findings support the need to automatically adapt the campaign definition, to maximize the relevant data collected.

References

1. Lin, J., Snow, R., Morgan, W.: Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In: Procs. of the 17th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, pp 422-429 (2011)
2. Plachouras, V., Stavrakas, Y.: Querying Term Associations and their Temporal Evolution in Social Data. In: Procs. of the 1st Intl. Workshop on Online Social Systems (2012)
3. Stavrakas, Y., Plachouras, V.: A Platform for Supporting Data Analytics on Twitter: Challenges and Objectives. In: Procs. of the 1st Intl. Workshop on Knowledge Extraction & Consolidation from Social Media (2012)
4. Allan, J. (ed.): Introduction to Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers (2002)
5. Dan, O., Feng, J., Davison, B.: Filtering microblogging messages for social tv. In: Procs. of the 20th Intl. Conf. companion on World Wide Web, pp. 197-200 (2011)
6. Ward, E.: Tweet Collect: short text message collection using automatic query expansion and classification. MSc thesis, University of Upsala (2013)
7. Ma, Z., Sun, A., Cong, G.: On Predicting the Popularity of Newly Emerging Hashtags in Twitter. *J. Am. Soc. Inf. Sci.*. doi: 10.1002/asi.22844
8. Tsur, O., Rappoport, A.: What's in a Hashtag? Content based Prediction of the Spread of Ideas in Microblogging Communities. In: Procs. of the 5th ACM Intl. Conf. on Web Search and Data Mining (2012)